To be presented at the European Conference on Machine Learning, ECML95, Heraklion, Greece:
Statistics, Machine Learning, and Knowledge Discovery in Databases Track, ECML95

AND to be presented at Daimler-Benz Research Center, Ulm, Germany

TITLE : Knowledge Discovery and Data mining: An overview
2-hour session.

Dr . Usama Fayyad
Technical Group Supervisor
Machine Learning Systems Group
Jet Propulsion Laboratory, California Institute of Technology
Pasadena, CA 911.09

Knowledge Discovery and Data Mining is the process of information extraction from very large databases. 'l'he extracted information is used either to explain phenomena in the mined data, or predict the outcomes based on new data. Database mining has recently received much attention, attested by the success of three Knowledge Discovery in Databases (KDD) AAAI workshops, and a book, because organizations realize that while they generate data in ever-increasing rates much of it is never exploited. The rapid growth of data and information created a need and an opportunity for extracting knowledge from databases, and both researchers and application developers have been responding to that need. KDD applications have been developed for astronomy, biology, finance, insurance, marketing, medicine, and many other fields. Core Problems in KDD include representation issues, search complexity, the use of prior knowledge, and statistical inference. The unifying themes include the use of domain knowledge, managing uncertainty, interactive (human-oriented) presentation, and applications.

The topics of interest covered in this presentation include:
Data preparation
% Data preprocessing techniques; filling missing values, data cleaning
% Selecting the data to be mined through database management, statistical analyses, and knowledge-based methods.
% Dimensionality reduction and feature space selection
Database mining techniques
% Artificial intelligence techniques (learning and conceptual clustering) .
% Neural networks (backpropagation, probabilistic neural network) .
% Case-based reasoning techniques (nearest neighbor anti its variants) .
% Deductive databases.
% Data visualization, and statistical analysis.
% Overview of Traditional approaches in Pattern Recognition

We will present the basic characteristics of each technique and illustrate each technique's applicability with examples from systems that have been presented in literature.

One of the hardest problems in database mining involves the selection & the most appropriate technique for extracting information from a particular data set. " This problem-is more acute when several techniques need to be used cooperatively. We will present criteria for selecting each of the discussed techniques based on the characteristics of the data and the analyst's goals. We will review four database mining systems developed by the authors for financial, manufacturing, Astronomy, and large image databases in planetary sciences.

The work described in this paper/presentation was carried out in part by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.